DOCUMENT RESUME

ED 464 962 TM 033 893

AUTHOR Lee, Yong-Won; Kantor, Robert; Mollaun, Pam
TITLE Score Dependability of the Writing and Speaking Sections of New TOEFL.
INSTITUTION Educational Testing Service, Princeton, NJ.
PUB DATE 2002-04-00
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002). For a related paper from the same authors, see TM 033 895.
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Ability; English (Second Language); *Generalizability Theory; Item Bias; Listening Comprehension Tests; *Reliability; *Scores; Second Language Learning; Speech; *Test Construction; Writing Tests
IDENTIFIERS *Test of English as a Foreign Language

ABSTRACT

This study examines the score dependability of writing and speaking assessments from the Test of English as a Foreign Language (TOEFL) from the perspectives of univariate and multivariate generalizability theory (G-theory) and presents the findings of three separate G-theory studies. For writing, the focus was on evaluating the impact on reliability of the number of raters (or ratings) per essay (one or two) and the number of tasks (one, two, or three). For speaking, the focus was on investigating the impact of the number of tasks (1 to 12) and the number of ratings (1 or 2). In addition, an attempt was made to determine the optimal configuration of types and number of tasks for speaking to maximize the reliability of the composite score. Data for the first study were ratings from a study of a new task type in which 2 reading-writing and 3 listening-writing tasks were administered to 488 English-as-a-Second-Language learners. For study 2, each of the essays for the tasks for one subgroup was rerated by each of six raters. In the third study, focusing on speaking, data from 261 examinees were used. By far the greatest source of variation in examinees' test performance was due to differences among test takers' ability as measured by the writing and speaking tasks. This suggests that, as intended, the tasks do distinguish among examinees. Results further suggest that, to maximize score reliability for both speaking and writing, it would be more cost efficient to increase the number of tasks rather than the number of ratings per task. Beyond five or six tasks, there would be a diminishing return for increasing the number of tasks. (Contains 5 tables, 6 figures, and 12 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 962

Score Dependability of the Writing and Speaking Sections of New TOEFL

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Yong-Won Lee
2A
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Yong-Won Lee

Robert Kantor

and

Pam Mollaun

Educational Testing Service

Rosedale Road Princeton

NJ 08541

Paper presented as a part of the symposium titled "Research in Support of the Development of New TOEFL" at the annual meeting of National Council on Measurement in Education (NCME) held between April 1-5 in New Orleans, LA.

Unpublished Work © 2002 by Educational Testing Service. All Rights Reserved.

TM033893

BEST COPY AVAILABLE

I. Introduction

If new types of writing and speaking tasks were to be included on a test, tasks and raters would be two potential sources of measurement errors. Both integrated and independent assessment tasks are possible candidates for assessing writing and speaking (Butler, Eignor, Jones, McNamara & Suomi, 2000; Cumming, Kantor, Powers, Santos & Talyor, 2000). These different types of tasks are associated with different types of input stimuli (e.g., auditory texts, reading passages, a stand-alone prompt). Potentially each of these different task types might measure a somewhat distinct aspect of speaking or writing, and reduce the reliability of a composite score based on a set of different task types. Another major factor that might impact score dependability (or reliability) might be raters (or ratings), since performance samples from examinees are rated by human raters. In particular, rating each writing or speaking sample only once is an attractive option because it would reduce costs. A critical question then is how much the score dependability would be impacted by combining scores on different task types into a composite score and rating each task only once.

Generalizability (G) theory provides a comprehensive conceptual framework and methodology for analyzing more than one measurement facet simultaneously in investigations of assessment error and score dependability (Brennan, 1992, 2000, 2001). In the classical test theory (CTT) framework, only one measurement facet can be examined at a time in relation to the object of measurement (persons), as exemplified by inter-rater or internal consistency reliabilities. The main purpose of the studies reported in this paper was to examine the combined impact on the score dependability of different testing conditions (*e.g.*, the number of tasks and raters, different rating designs) for the new task types being evaluated.

This document examines the score dependability of writing and speaking assessment from the univariate and multivariate G-theory perspectives and presents the findings from the three separate G-*theory* studies. For writing, a special focus was placed on evaluating the impact on the reliability of the number of raters (or ratings) per essay (1 or 2) and the number of tasks (1, 2, or 3). For speaking, the focus was on investigating the impact of the number of tasks (varied from 1 to 12) and of the number of ratings (1 or 2). In addition, an attempt was made to determine the optimal configuration of types and number of tasks for speaking to maximize the reliability of the composite score.

II. Score Dependability of Writing and Speaking Measures in the G-theory framework

G- theory (Brennan, 1992; Brennan, 2000) is a comprehensive framework that enables us to disentangle multiple sources of error from the test data, instead of putting them all into a single undifferentiated error term, as in the CTT. Through the *generalizability* (G) and *decision* (D) studies, we can evaluate the relative importance of various sources of measurement error, investigate the impact of various changes in the measurement design (*e.g.*, different numbers of tasks or raters, standardized rating procedures), and interpret the test score reliability both from norm-referenced and

criterion-referenced frameworks (*e.g.*, a generalizability coefficient (G) and a dependability index (Φ)).

G-theory can be also extended for the multivariate situations where we have a test made up of multiple subsections or subtests and need to investigate the reliability of the composite of subsection scores (Brennan, 2001). Both univariate and multivariate G-theories have been developed within the same G-theory framework, but differ in that the former is suited to analyzing scores from a single test at a time, whereas the latter can also handle scores from multiple subtests simultaneously. For instance, differences between the univariate and multivariate G-theory analyses can be also illustrated in a test developed according to the same "table of specifications". In this type of tests, the same structure of content subsections is applied for assembling multiple forms of the same test. If we are simply interested in the impact on the score reliability of the different numbers of items in the whole test, ignoring the subsection boundary, we can just use the univariate G-theory to estimate variance components and score reliability coefficients for the total test. However, if we are interested in what combination of subtest lengths (or weights) for a fixed total test length can maximize or minimize the composite score reliability of the test, the univariate G-theory analysis does not provide answers to such inquiries, but the multivariate G-theory analysis can.

In the multivariate G- theory design, test subsections or content categories (v) are considered a fixed facet, and the number of elements in each fixed category can be different or unbalanced. It is possible to estimate variance components for each fixed content category and compute the G and Φ coefficients of the composite scores within the framework of multivariate G-theory.

Rating Designs for G- and D- studies.

In the generalizability stage of a G-study, the variances associated with various facets of measurement including the object of measurement are estimated (Suen, 1990; Lynch & McNamara, 1998). When multiple raters (r) and tasks (t) are involved in the assessment of examinee (p) proficiency, the most powerful G-study designs might be a fully-crossed design with raters crossed with tasks and examinees (p x t x r). Once the data are collected according to such a fully-crossed design, it will allow the researchers to investigate the score reliability for various possible rating designs involving persons, raters, and tasks. However, this design is rarely feasible for large-scale testing situations, except for a specially-designed rating session with a smaller sample of essays for re-rating.

The next most probable rating design in an operational testing situation would be a partially-nested design, with raters nested within examinees ((r:p) x t). The G-theory analyses based on the nested design, however, has not been available for the TOEFL family tests, because such a nested design was not used in rating the performance samples for operational tests in a strictest sense. Although two raters rate the performance samples for each examinee, the same rater pair is not systematically assigned to a block of examinees, as in a typical nested design. Instead, the rater pair is

somewhat randomly decided for each examinee. The same has been true for the pilot tests of new writing and speaking tasks used in this study.

One practical solution in such circumstances is to treat ratings (*i.e.*, the first and second ratings: r') as a random facet instead of raters (r). Since all examinees' final scores are based on the two ratings, it is possible to use a fully crossed design with two random facets of "tasks" and "ratings" (not raters) (*p* x t x r'). This alternative strategy is also consistent with the inter-rater reliability computation procedure used for large-scale performance-type language assessments, such as the Test of Spoken English (TSE), the Test of Written English (TWE), and the TOEFL CBT essay test. The ratings (not raters) are to be used as a unit of analyses to obtain the inter-rater reliability estimate. The same approach has been adopted by a number of researchers in language assessment as an alternative (Bachman, Lynch, & Mason, 1995; Lee & Carey, 2001; Lee, Golub-Smith, Payton & Carey, 2001). For this reason, the *p* x t x r' design was also initially used for the new writing and speaking tasks used in this study.

In the decision (D) study stage, consequences of various changes in the measurement design are investigated based on the estimated variance components for the main and interaction effects for the measurement facets involved in the assessment system. For instance, it is possible to investigate the impact of using only one rating per essay for different test lengths instead of using the mean score across two raters. By re-rating writing samples for a sub-sample of examinees according to a strictly-crossed (p x t x r) design, we may be also able to investigate the impact on the score dependability of using more differentiated one-rating-per-essay scenarios (*e.g.*, using a single rating per an essay but having each task for the same test-taker rated by different raters) than in the p x t x r' design. In the multivariate design, the different combinations of subsection length for a fixed number of total test lengths can be compared in terms of G- and Φ coefficients.

Generalizability coefficient and dependability index

Univariate analysis. In the p x t x r' design for the univariate analysis, all of the 7 main and interaction variance components (σ_p^2, σ_t^2, $\sigma_{r'}^2$, $\sigma_{pr'}^2$, σ_{pt}^2, $\sigma_{tr'}^2$, $\sigma_{ptr'}^2$) can be estimated for a single observation in the G-study. In the D-study, two different kinds of score reliability equivalents can be computed for different measurement scenarios, that is, a *generalizability* coefficient (G or $E\rho^2$) and a *dependability* index (Φ).

First, the relative error term (δ) and the *generalizability* coefficient (G) can be defined, as in equation 1, which can be interpreted as the error variance and the reliability index for norm-referenced score interpretation, respectively (Brennan, 1992; Suen, 1990).

$$G = \frac{\sigma_p^2}{\sigma_p^2 + \delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{pr'}^2}{n_{r'}} + \frac{\sigma_{ptr'}^2}{n_t n_{r'}}} \quad \text{.....(1)}$$

Second, the absolute error (Δ) and the dependability index (Φ) can be computed, as in equation 2, which can be interpreted as the error variance and the score reliability index for criterion-referenced score interpretation, respectively. When the mean score is applied as a cut-off score for admission and licensure purposes, the Φ coefficient is more appropriate.

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \Delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_t^2}{n_t} + \frac{\sigma_{r^\bullet}^2}{n_{r^\bullet}} + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{pr^\bullet}^2}{n_{r^\bullet}} + \frac{\sigma_{tr^\bullet}^2}{n_t n_{r^\bullet}} + \frac{\sigma_{ptr^\bullet}^2}{n_t n_{r^\bullet}}} \quad \text{.....(2)}$$

Multivariate Analysis. In the univariate design, it is not possible to examine the score reliabilities for different combination of subsection length for a fixed section length, because content subcategories are a hidden facet. For instance, several different combinations of three subsection lengths would be possible for a total section length of 5 tasks for a speaking section. An attractive option for this situation is to recognize modality subsections as a fixed facet in multivariate $p^\bullet \times t^\circ \times r^{\bullet\bullet}$ design. In the multivariate design, a superscript filled-circle ($\bullet$) next to a facet symbol indicates that the facet is crossed with the fixed category facet (v), whereas a superscript empty circle ($^\circ$) signals that the facet is nested within the multivariate variables (v).

In the $p^\bullet \times t^\circ \times r^{\bullet\bullet}$ design for multivariate analysis, all of the 7 main and interaction variance components (σ_p^2, σ_t^2, $\sigma_{r^\bullet}^2$, $\sigma_{pr^\bullet}^2$, σ_{pt}^2, $\sigma_{tr^\bullet}^2$, $\sigma_{ptr^\bullet}^2$) are estimated for each of the levels for the fixed content category facet (v).

First, the relative error term for the composite score ($\sigma_C^2(\delta)$) and the composite score generalizability coefficient (G or $E\rho^2$) can be defined, as in equation 3, which can be interpreted as the error variance and the reliability index for norm-referenced score interpretation, respectively (Brennan, 1992).

$$G = \frac{\sigma_C^2(\tau)}{\sigma_C^2(\tau) + \sigma_C^2(\delta)} = \frac{\sum_v \sum_{v'} \omega_v \omega_{v'} \sigma_{vv'}(\tau)}{\sum_v \sum_{v'} \omega_v \omega_{v'} [\sigma_{vv'}(\tau) + \sigma_{vv'}(\delta)]} \quad \text{.....(3)}$$

Second, the absolute error for the composite score ($\sigma_C^2(\delta)$) and the composite score dependability index (Φ) can be computed, as in equation 4, which can be interpreted as the error variance and the score reliability index for criterion-referenced score interpretation, respectively.

$$\Phi = \frac{\sigma_C^2(\tau)}{\sigma_C^2(\tau) + \sigma_C^2(\Delta)} = \frac{\sum_v \sum_{v'} \omega_v \omega_{v'} \sigma_{vv'}(\tau)}{\sum_v \sum_{v'} \omega_v \omega_{v'} [\sigma_{vv'}(\tau) + \sigma_{vv'}(\Delta)]} \quad \text{.....(4)}$$

III. Methods

Data:

Writing. Data used for study 1 are ratings from a study of new task types. In this study, two Reading-Writing and three Listening-Writing tasks were administered to 488 ESL examinees. Examinee responses that were rated on a scale of 1 to 5 by two raters from a pool of 27 raters. There were 3 independent (TWE-like) writing tasks used, but each of them was given to three different subgroups of examinees (n_{g1}=162, n_{g2}= 164, n_{g3} =162). Thus each subgroup of examinees completed 6 writing tasks (*i.e.*, 5 common integrated tasks +1 independent task specific to each subgroup).

For Study 2, each of the essays for the 6 tasks for subgroup 3 (n_{g3} =162) was re-rated by each of the 6 raters on a scale of 1-5. Data were collected according to a fully-crossed design (p x t x r). In other words, all 6 writing samples from each of the 162 examinees were rated by all 6 six raters. Based on the examination of performance of the raters who had participated in the previous rating session, the six most typical raters were chosen for this study.

Speaking. Data used for this study are also from the study of new task types. Six Listening-Speaking (LS), two Reading-Speaking (RS), and five Independent Speaking (IS) tasks were administered to the examinees. The examinee responses were rated on a scale of 1 to 5 by two independent raters from a pool of 22 raters. Data from only 261 examinees were used for the analysis. Only those examinees who had scores for at least two tasks of all three types were included in the analysis. In addition, two Listening-Speaking tasks based on academic conversations were excluded, because it was determined that they did not provide enough substance for spoken responses. Therefore, there were only 11 tasks (4 LS, 2RS, 5IS) included in this study.

Analysis:

The computer program GENOVA (Crick & Brennan, 1983) was used to estimate the variance components for the main and interaction effects and G and Φ coefficients for the univariate analyses for both the writing tasks (Studies 1 and 2) and the speaking tasks (Study 3). The computer program mGENOVA (Brennan, 1999) was used to conduct the multivariate G-theory analysis for speaking.

Writing (Study 1). For Study 1, separate analyses were done to estimate the variance components for each of the three subgroups of examinees (n_{g1}=162, n_{g2}= 164, n_{g3} =162) based on the initial data set. The three subgroups were treated as if each of the examinee subgroups had taken 3 different forms of the same test, with LW and RW tasks as common items across all three forms. Each of the variance components was averaged across the three subgroups to obtain stable estimates, and then the G- and Φ coefficients were computed based on these more stable, averaged variance components. Since there was only one task in the IW subsection, it was not possible to do a multivariate G-theory

analysis with all three subsections included in a fixed content facet (v) for writing. There should be at least two tasks in each subsection to estimate variance associated with tasks in the G-study for the subsections.

In the G-study, a two-facet crossed design (p x t x r') with tasks and ratings as random facets was employed to estimate the variance components for the main and interaction effects. Based on the estimates, multiple D-studies for computing G and Φ coefficients were carried out by varying the number of tasks from 1 to 6 and the number of ratings from 1 to 2. On the other hand, a single facet (p x t) design was used for the averaged rating over two raters to estimate internal consistency reliabilities (α) for different section lengths. Cronbach's (1951) coefficient α equals a *generalizability* (G) coefficient in a single-facet crossed design (Brennan, 1992; Suen, 1990). Finally, internal consistency reliabilities from the p x t design were plotted with the G and Φ coefficients that were estimated from the p x t x r' design for different testing conditions.

Writing (Study 2), a two-facet crossed design (p x t x r) with tasks and raters as random facets was used to estimate the variance components in the G-Study based on the re-rated data for subgroup 3 (n_p=162). Estimated variance components were compared to those from the p x t x r' design from the original data. In the D-study, the G- and Φ coefficients were estimated for different numbers of tasks and raters not only for the R:(p x T) but also other related designs of interest, such as p x T x R, (R:p) x T. The R:(p x T) design represents the rating scenario of a particular interest in which each task for a single examinee is rated by different raters. The G and Φ coefficients estimated from these designs were compared to those from the p x t x r' design based on the original data for subgroup 3.

Speaking. For the *univariate analysis*, the same two-facet crossed design (p x t x r') as used in Study 1 for writing was employed for speaking as well to estimate the variance components in the G-study. Multiple D-studies were carried out by varying the number of tasks from 1 to 12 and the number of ratings from 1 to 2. A single facet (p x t) design was used for the averaged rating over two raters to estimate the internal consistency reliability for the different test lengths.

For the multivariate analysis, a two-facet crossed design ($p^{\bullet}$ x t^{O} x r'$^{\bullet}$) in the G-study to estimate the variance components for the main and interaction effects for each subsection of task types. It was assumed that persons (p) and ratings (r') were crossed with the subsections (v) of LS, RS, and IS, but tasks (t) were nested within each subsection (v). Multiple D-studies were carried out by varying the number of tasks in each subsection for the several fixed section lengths. Of particular interest were comparisons of composite score reliabilities for different combinations of subsection lengths for the fixed total test lengths of 3, 4, and 5 tasks. When the total section length was 3, the only possible scenario for representing all three subsections was to take one task for each of the LS, RS, and IS subsections (1-1-1). For the test length of 4, there were three possible scenarios (1-1-2, 1-2-1, and 2-1-1). When the total test length was 5, there were 6 possible scenarios, which included 1-1-3, 1-2-2, 1-3-1, 2-2-1, 3-1-1, 2-1-2. For comparison purposes, two additional combinations for the total test length of 6 were

included in the D-study, along with the two longer test scenarios that were used in the G-study to estimate the variance components (4-2-5, 4-3-5).

IV. Results and Discussion

For both writing and speaking, similar results were obtained about the relative effects of variance components related to tasks and ratings that were estimated through the univariate G-theory analyses. We found that adopting a single rating scheme had a smaller effect on the score dependability than expected. One of the striking differences between writing and speaking was that the variance components associated with the task main effect (t) for speaking is much smaller in proportion (about a fifth) than in writing. The univariate analysis on re-rated data (n=162) for writing has also supported the findings from the original data. Finally, the multivariate G- theory analysis on speaking has provided very useful information in evaluating several possible final configurations of the speaking tasks in terms of maximizing the dependability of the composite speaking scores.

Writing (Study 1)

Separate analyses based on three subgroups of examinees and one based on the averaged estimates have yielded similar patterns about the relative proportion of each of the variance components and score reliabilities for various testing conditions. Since an assessment design of particular interest in this study was a test of 3 writing tasks taken by all examinees and rated once or twice, the results are discussed with a focus on these scenarios.

First, the variance component associated with examinees (p) was the largest, which explained about 39% of the total variance estimated for a single observation in the G-study, as shown in Table 1-a. In the single rating situation for 3 tasks, about 65% of the total variance was explained by the person (p) variance, whereas 46% was explained in the double rating situation for a single task. The main effect for ratings (R') was zero, indicating that the ratings were not different in severity between the first and second ratings. However, the main effect for tasks (T) accounted for about 11% of the total variance in the G-study, which may mean that there is a considerable difference in task difficulty among writing tasks.

The examinee-by-task-by-rating interaction (pTR') accounted for the second largest proportion of the total variance for a single observation in the G-study. The pTR' component explained about 30% of the total variance for a single observation in the G-study, but this proportion was reduced to 16% for the 3 task scenario with a single rating per essay. The next largest interaction component was the person-by-task interaction (pT) variance component (*i.e.*, 20% for a single observation, 11% of the total variance for the 3-task-and-single-rating scenario). This interaction may indicate some sub-groupings of either tasks or examinees or task-rating combinations that are behaving differently from the rest of the groups. Nevertheless, the examinee-by-rating interaction variance (pR') component was very small (0.5%), and the task-by-rating (TR') was almost zero.

Second, due to the fact that the examinee-by-rating interaction (*p*R') variance was negligible, the impact of one rating per essay was not as serious as expected, as shown in Table 1-b and Figure 1. For instance, the G-coefficient was reduced from about .77 to .69, when the number of ratings was decreased from 2 to 1 for the 3-task situation. Reductions of .08 to .09 in the G-coefficient resulted consistently when the number of ratings was reduced to 1 for each of the 1, 2, and 3 task situations. As the test length increased, the reduction in the G-coefficient value became smaller. However, when the number of tasks was increased from 1 to 3 for the single rating scenario, there was a dramatic .25 increase in the G coefficient (*i.e.*, .17 from 1 to 2, .08 from 2 to 3). The similar pattern was observed for the Φ coefficients. It seems that the lost portion of score reliability due to the adoption of the single-rating-per-essay scheme could be probably recovered more than enough by increasing the number of tasks from 1 to 3 for the writing.

Finally, as shown in Figure 1, the line for internal consistency alpha (α) was very close or almost identical to that for the G coefficient for the two-rating-per-essay scenario. This may be attributed to the fact that variance components associated with both the rating main (R') and the person-by-rating interaction (*p*R') effects were very small, close to zero in the data. Usually, the coefficient α from the p x t design based on the averaged rating over two raters should be bigger than the G-coefficient estimated from the p x t x r' design for the two-rating situation, because the random rating error was not incorporated in the α estimate.

Writing (Study 2)

Similar patterns were observed for the relative proportions of task- and rating-related variance components in the original (n_p=162, n_t = 6, $n_{r'}$ = 2) and the re-rated data (n_p=162, n_t = 6, $n_{r'}$ = 6) for subgroup 3. A prime motivation for re-rating the essays for subgroup 3 was to investigate whether or not the error associated with rater judgment was underestimated and, consequently, the score reliability was overestimated in the p x t x r' design based on the original data. As expected, it was found that the rating-related error variance components (r', pr') estimated from the original data were slight underestimates for the relative effects of rater-related error (r, pr). On the other hand, the relative proportions of task-related error (t, pt) seemed to have been slightly overestimated in the p x t x r' design than in the p x t x r design. Most importantly, the effect of the examinee-by-task-by-rating interaction (ptr') from the p x t x r' design would be a significant overestimate for the three-way interaction component (ptr) from the p x t x r design. All these taken together, the G- and Φ coefficients from the original data seemed to be slight underestimates (rather than overestimates as concerned) of the score reliability for the writing section computed from the complete p x t x r design.

First, the variance component associated with examinees (*p*) was the largest in both the original and the re-rated data, but it accounted for a greater proportion of the total variance in the re-rated than in the original data. In the re-rated data, for instance, it explained about 49% of the total variance estimated for a single observation in the G-study, but only 41% in the original data for subgroup 3 (39% on average across 3

subgroups), as shown in Tables 2-a and 2-b. For the 3-task-and-single-rating situation, about 77% of the total variance was explained by the person (p) variance in the re-rated data, whereas only 67% was explained in the original data. The main effect for raters (R) occupied about 1.5% of the total score variance in the re-rated data, while the rating main effect (R') variance was zero in the original data. However, the main effect for tasks (T) accounted for a slightly smaller portion of the score variance in the re-rated (8%) than in the original data (11%) in the G-study.

The examinee-by-task-by-rater interaction (pTR) was the second largest among the variance components in the re-rated data, but its relative proportion in the total score variance was significantly smaller than that of its three-way interaction counterpart (pTR') in the original data. The pTR effect explained about 20% of the total variance in the re-rated data, but the pTR' explained as much as about 30% in the original data. When the number of tasks was increased from 1 to 3 tasks, the relative effect of the pTR variance was dramatically reduced to 8% in the D-study for a single rating situation (See Table 2-b). The next largest interaction variance component was the person-by-task interaction (pT) variance accounting for about 18% of the total variance in the G-study in the original and re-rated data sets. The relative proportion of the pT variance was also reduced to 7%, when the number of tasks was increased to 3. Finally, both the examinee-by-rater (pR) and the task-by-rater (TR) interaction variance components were very small, explaining only 1.2% and 1.6% of the total score variance in the re-rated data, respectively.

Second, the impact on the G-coefficient of decreasing the number of raters from 2 to 1 was not as big as that of increasing the number of tasks from 1 to 3 for the three D-study designs based on the re-rated data, as shown in Tables 3-a and 3-b. For the R:(p x t) design, for instance, reductions of approximately .04 to .06 in the G-coefficient were observed when the number of ratings was reduced to 1 for each of the one-, two-, and three- tasks situations. As the test length increased, the reduction became smaller. However, when the number of tasks was increased from 1 to 3, with the number of ratings per essay fixed, there was a dramatic .24 increase in the G coefficient (*i.e.*, .17 from 1 to 2, .07 from 2 to 3). A similar pattern was observed for the Φ coefficients. It seems apparent that the lost portion of score reliability due to the reduced number of ratings could be probably compensated for more than enough by increasing the number of tasks to three even in the R:(p x t) design, which has basically confirmed the findings from Study 1 based on the p x t x r' design.

Finally, among the three designs adopted for the re-rated data in this study (i.e., p x T x R, (R:p) x t, and R:(p x t)), it seemed that the p x t x r design would be maximizing the G coefficient for the 3-task and single-rating situation, while the R:(p x t)) and (R:p) x t designs, the Φ coefficient. Nevertheless, the G- and Φ coefficient differences among the three designs were not very large, as shown in Figures 2-a and 2-b. In addition, the G- and Φ coefficients estimated from the p x t x r' design were consistently lower than those from the R:(p x t)) based on the re-rated data, as shown in Figure 3. This suggests that we did not overestimate the score reliability for writing in Study 1 based on the original data.

Speaking (Study 3)

Similar results were obtained about the impact of single rating on the score reliability both from the univariate and multivariate analyses. Multivariate analysis provided additional useful information for comparing various scenarios for the fixed total section lengths. Assessment designs of particular interest were efficient scenarios of 5 speaking tasks taken by all the examinees and rated once or twice in comparison to a liberal scenario of 12 tasks rated twice. The results of the analyses are discussed with a focus on these scenarios.

Univariate Analysis. First, the variance component associated with examinees (p) was the largest, and explained about 51 % of the total variance estimated for a single observation in the G-study, as shown in Table 4. In the D-study, about 82% of the total variance was explained by the person (p) variance for the 5-task-and-single-rating scenario, whereas about 93% was explained for the 12-task-and-double-rating situation. The main effect for ratings (R') was zero, indicating that the ratings were not different in overall severity between the first and second ratings. However, the main effect for tasks (T) accounted for about 2% of the total variance in the G-study, which may mean that there is only a small difference in task difficulty among tasks, especially compared to writing.

The examinee-by-task-by-rating interaction (pTR') accounted for the second largest proportion of the total variance for a single observation in the G-study (*i.e.*, about 28% for a single observation). The relative proportion in the total variance of the pTR' component was drastically reduced from 28% to 9%, when the number of tasks was increased to 5 for the single rating scenario. Nevertheless, when the numbers of tasks and ratings were further increased to 12 and 2, respectively, this proportion has decreased to 2%, which could be seen as a small change for adding seven more tasks and one more rating per essay. The next largest interaction component was the person-by-task interaction (pT) variance component (*i.e.*, 17% for a single observation, 6% and 3% of the total variance for the above two D-study scenarios). However, the relative effect of the examinee-by-rating interaction variance (pR') component was very small (1.8%), and the task-by-rating (TR') variance was almost zero.

Second, due to the fact that the examinee-by-rating interaction (pR') variance was small, the impact of one rating per essay was not as serious as expected, as shown in Table 4-b and Figure 5. For instance, the G-coefficient was reduced from about .88 to .83 (.05 reduction) for the 5-speaking-task test, when the number of ratings was decreased from 2 to 1. Reduction in G-coefficient values due to the single rating ranged from .04 to .09 for various section lengths of 1 to 12 tasks. However, when the number of tasks was increased from 1 to 5 for the single-rating scenario, the G-coefficient was increased from .52 to .83 (.31 increase). There was a .17 additional increase in the G-coefficient (from .83 to .90), when the numbers of tasks was further increased from 5 to 12. A similar pattern of impact was observed for the Φ coefficients.

Third, as shown in Figure 5, the internal consistency alpha (α) was slightly higher than the G coefficient for the two-rating-per-essay scenario, but very close to the coefficient. This may be attributed to the fact that variance components associated with both the rating main (R') and the person-and-rating interaction effects (pR') were comparatively small for speaking as they were for writing.

Multivariate Analysis. First, the person (p) variance was the largest in each of the three subsections, explaining about 56.3%, 55.7%, and 53.4% for the LS, RS, and IS subsections, respectively (See Table 5). The main effect for ratings (r') was close to zero in the G-study. On the other hand, the main effect for tasks (t) accounts for about 1.4%, 0.4%, and 0.3% of each of the total score variances for each subsection in the G-study, which may mean that there is a very small difference in difficulty across tasks in each subsection. In all three subsections, the second largest variance component was the examinee-task-and-rating interaction (pTR') in the G-study. In the LS and IS subsection, it was followed by the examinee-task (pT) interaction variance in size, but, in the RS subsection, the examinee-by-rating variance (pR') was the third largest variance component followed by the pT variance. Especially, the pR' variance component was significantly smaller in the LS subsection than in other two subsections, which was partially responsible for the error variance being the smallest and the universe score variance (p) being the largest proportionally in the LS subsection.

Second, the correlations between LS and RS subsection scores (.98) and between RS and IS subsection scores (.95) were very high, whereas that between the LS and IS subsection scores was comparatively lower (.89). The very high correlation between the two integrated task types may suggest that tasks in these two subsections may be measuring a very similar construct, even though the stimuli input modes are different. On the other hand, the correlation between the IS and LS subsection scores the lowest among the three subsection pairs (though it can still be seen as high), probably because two task types are comparatively more dissimilar than other pairs in terms of both stimuli input modes (auditory vs. visual) and the nature of tasks (integrated vs. independent).

Third, because the examinee-by-rating interaction (pR') variance was small, the impact of using one rating per task was less serious than expected, as shown in Table 5-b and Figure 6. For instance, the G-coefficient for the composite scores was reduced from about .85 to .78 (.07 reduction) for the 3-speaking task test with one task in each subsection, when the number of ratings was decreased from 2 to 1. For other configurations of longer section lengths, the one-rating-per-speech-sample has reduced G-coefficients by .04 to .06. A similar pattern of the impact was observed for the Φ coefficients for the composite scores.

Fourth, it seemed that there would be largest gains in composite score reliability, when the number of LS tasks was increased, as shown in Figure 6, because there was slightly less psychometric noise among tasks in the LS subsection. Among the six scenarios for the fixed section length of 5 tasks, the scenario of three LS, one RS, and one IS tasks (3-1-1) seems to achieve the highest G- and Φ coefficients for both one and two rating situations. Similarly for the test length of 4, the highest G- and Φ coefficients were

obtained for the 2-1-1 scenario. However, the actual differences in score reliability values among different combinations of subsection lengths for a fixed section length were not seriously large.

V. Conclusion

By far the greatest source of variation in examinees' test performances was due to differences among test takers' ability as measured by the writing and speaking tasks. This suggests that, as intended, the tasks do distinguish among examinees. Also, as has been found in numerous other generalizability studies, the next greatest proportion of test variance was attributable to the interaction of person-by-task and person-by-task-by-rater. The variance due to tasks was minimal (particularly for speaking). These latter two findings suggest that, while tasks are on average comparable in difficulty, they are not *uniformly* difficult for all examinees. Finally, the variance due to ratings (or raters) was close to zero, and the examinee-by-rating (rater) was also very small for both writing and speaking.

Despite these concerns raised about using rating (r') as a random facet for writing and speaking, however, the analyses of re-rated writing data have shown that the p x t x r' design resulted in similar (probably more conservative) estimates of score reliabilities.

These results indicate that, to maximize score reliability for both speaking and writing, it would be more cost-efficient to increase the number of tasks rather than number of ratings per task. However, beyond 5 or 6 tasks, there would be clearly a diminishing return for increasing the number of tasks.

Among the three subsections for speaking, performance among examinees was slightly more uniform among tasks in the LS (Listening-Speaking) subsection, and, thus, it appears that more gains in composite score reliability could result from increasing the number of LS tasks for the fixed section lengths. However, the actual differences in score reliability values among different combinations of the number of tasks in each subsection for the fixed total section length were quite small. Therefore, the decisions about the possible test configurations should be based on the content and test development considerations as well as the score reliabilities.

V. References

Bachman, L. F., Lynch, B. K. & Mason, M. (1995). Investigating variability in tasks and rater judgment in a performance test of foreign speaking. *Language Testing, 12*, 238-257.

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: ACT.

Brennan, R. L. (1999). *Manual for mGENOVA Version 2.0.* Iowa City, IA: The University of Iowa.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24 (4)*, 339-353.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL monograph MS -20). Princeton, NJ: ETS.

Cumming, A., Kantor, R., Powers, D., & Santos, T. (2000). TOEFL 2000 writing framework: A working paper (TOEFL monograph MS -19). Princeton, NJ: ETS.

Crick, G. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system (ACT Tech. Bulletin No. 43).* Iowa City, IA: American College Testing Program.

Lee, Y. (April, 2001). *The essay scoring and scorer reliability in TOEFL CBT.* Paper presented at a Computer-based TOEFL symposium at the annual meeting of National Council on Measurement in Education (NCME) in Seattle, WA.

Lee, Y., Golub-Smith, M., Payton, C., & Carey, J. (April, 2001). *The score reliability of the Test of Spoken English (TSE) from the generalizability theory perspective: Validating the current procedure.* Paper presented at the annual conference of American Educational Research Association (AERA) in Seattle, WA.

Lynch, B. K. & McNamara, T. F. (1998). Using g-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing 15* (2), 158-80.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum.

Table 1. Estimated variance components averaged across three examinee subgroups and G- and Phi coefficients for Writing (n_{total} =488; n_1=162, n_2=164, n_3=162)

(a) Estimated Variance Components

Effects	G-Study (p x t x r')		D-Study (*p* x T x R')			
	Single Observation		1 Task, 2 Rating		3 Tasks, 1 Rating	
	Variance	Percent	Variance	Percent	Variance	Percent
Person (*p*)	0.50568	39.07%	0.5057	45.99%	0.5057	65.43%
Task (T)	0.13702	10.59%	0.137	12.46%	0.0457	5.91%
Rating (R')	0	0.00%	0	0.00%	0	0.00%
Person-by-Task (*p*T)	0.26213	20.25%	0.2621	23.84%	0.0874	11.31%
Person-by-Rating (*p*R')	0.00645	0.50%	0.0032	0.29%	0.0065	0.83%
Task-by-Rating (TR')	0.00077	0.06%	0.0004	0.03%	0.0003	0.03%
Person-by-Task-By Rating (*p*TR')	0.3822	29.53%	0.1911	17.38%	0.1274	16.48%
Total	1.29424	100.00%	1.0995	100.00%	0.7728	100.00%

(b) Estimated G- and Φ coefficients based variance components averaged across 3 examinee subgroups

No. of Tasks	*p* x T x R' Design (Based on Averaged Variance)				
	One Rating Per Essay		Two Ratings Per Essay		Two Ratings Per Essay
	G	Φ	G	Φ	α
1	0.43669	0.39029	0.525	0.45935	0.5284
2	0.60532	0.55914	0.6867	0.62793	0.6913
3	0.69479	0.65344	0.7654	0.71556	0.7705
4	0.75026	0.71365	0.812	0.76927	0.8174
5	0.78801	0.75543	0.8427	0.80557	0.8484
6	0.81536	0.78611	0.8645	0.83174	0.8703

Figure 1. Internal consistency reliabilities (α) for averaged ratings and generalizability coefficients (G or $E\rho^2$) and dependability indices (Φ) for one- and two rating per essay scenarios averaged over 3 subgroups.

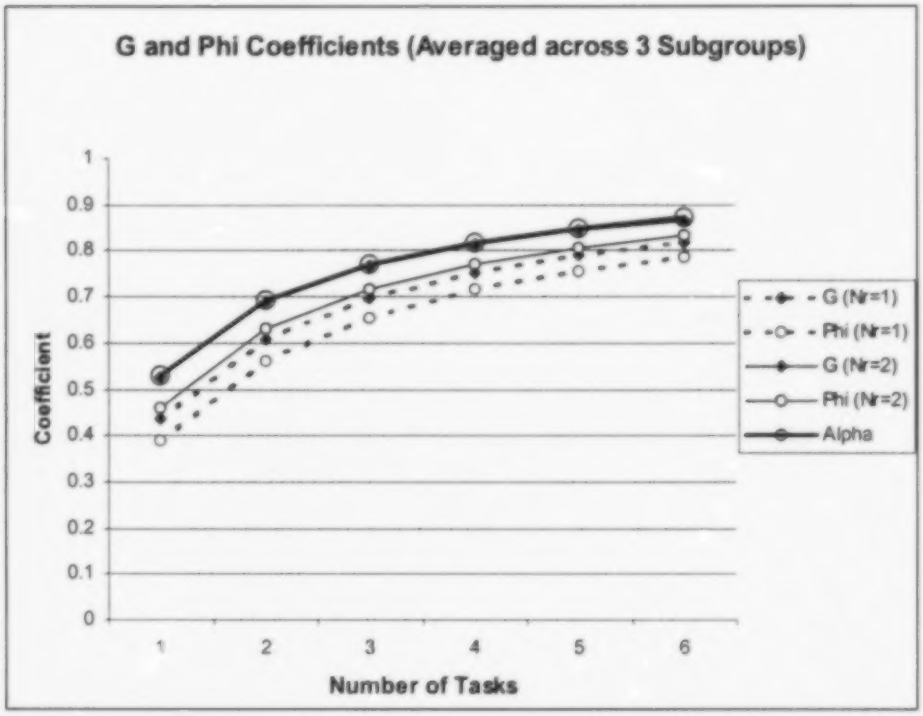


Table 2. Estimated Variance components from the G- and D studies from the univariate analyses of the initial and re-rated data for a subgroup of examinees (n=162) for Writing

(a) Initial Data (n_p=162, n_t = 6, $n_{r'}$ = 2)

Effects		G-Study (p x t x r')		D-Study (*p* x T x R')			
		Single Observation		1 Task, 2 Rating		3 Tasks, 1 Rating	
		Variance	Percent	Variance	Percent	Variance	Percent
Person	(*p*)	0.54494	40.86%	0.5449	48.17%	0.5449	67.45%
Task	(T)	0.14071	10.55%	0.1407	12.44%	0.0469	5.81%
Rating	(R')	0	0.00%	0	0.00%	0	0.00%
Person-by-Task	(*p*T)	0.24324	18.24%	0.2432	21.50%	0.0811	10.04%
Person-by-Rating	(*p*R')	0	0.00%	0	0.00%	0	0.00%
Task-by-Rating	(TR')	0.0023	0.17%	0.0012	0.10%	0.0008	0.10%
Person-by-Task-by Rating	(*p*TR')	0.40264	30.19%	0.2013	17.79%	0.1342	16.61%
Total		1.33383	100.00%	1.1314	100.00%	0.8079	100.00%

(b) R-rated Data (n_p=162, n_t = 6, n_r = 6)

Effects		G-Study (p x t x r)		D-Study (*p* x T x R)			
		Single Observation		1 Task, 2 Rater		3 Tasks, 1 Rater	
		Variance	Percent	Variance	Percent	Variance	Percent
Person	(*p*)	0.70665	48.93%	0.70665	55.83%	0.70665	76.84%
Task	(T)	0.12164	8.42%	0.12164	9.61%	0.03041	3.31%
Rater	(R)	0.02115	1.46%	0.01057	0.84%	0.02115	2.30%
Person-by-Task	(*p*T)	0.25899	17.93%	0.25899	20.46%	0.06475	7.04%
Person-by-Rater	(*p*R)	0.01696	1.17%	0.00848	0.67%	0.01696	1.84%
Task-by-Rater	(TR)	0.02287	1.58%	0.01143	0.90%	0.00572	0.62%
Person-by-Task-by Rater	(*p*TR)	0.29592	20.49%	0.14796	11.69%	0.07398	8.04%
Total		1.44418	100.00%	1.26572	100.00%	0.91962	100.00%

Table 3. Estimated G and Φ coefficients for the four D-study designs based on the original and re-rated data for subgroup 3.

(a) *p* x T x R' (original) vs. *p* x T x R (re-rated)

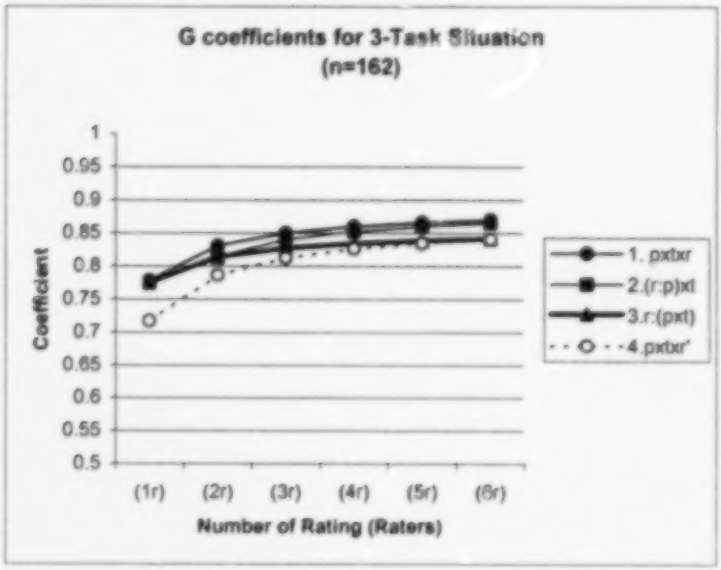
No. of Tasks	p x T x R' Design				p x T x R Design			
	One Rating Per Essay		Two Ratings Per Essay		One Rating Per Essay		Two Ratings Per Essay	
	G	Φ	G	Φ	G	Φ	G	Φ
1	0.45762	0.40855	0.55072	0.48167	0.55271	0.48931	0.62977	0.5583
2	0.6279	0.5801	0.71028	0.65017	0.7059	0.64566	0.76926	0.70969
3	0.71681	0.67451	0.78621	0.73599	0.77775	0.72262	0.83059	0.78022
4	0.77142	0.73426	0.8306	0.788	0.81946	0.76842	0.86507	0.82101
5	0.80838	0.77548	0.85973	0.82289	0.8467	0.7988	0.88717	0.8476
6	0.83505	0.80562	0.88031	0.84792	0.86589	0.82042	0.90254	0.8663

(b) R:(*p*xT) and (R:*p*) x T designs (re-rated)

No. of Tasks	R:(*p*xT) Design				((R:*p*) x T) Design			
	One Rating Per Essay		Two Ratings Per Essay		One Rating Per Essay		Two Ratings Per Essay	
	G	Φ	G	Φ	G	Φ	G	Φ
1	0.53432	0.48931	0.59231	0.53751	0.53432	0.48931	0.61765	0.5583
2	0.69649	0.6571	0.74396	0.6992	0.68365	0.64566	0.75586	0.70969
3	0.77488	0.7419	0.81338	0.77712	0.75388	0.72262	0.81678	0.78022
4	0.82109	0.79307	0.85319	0.82297	0.7947	0.76842	0.85108	0.82101
5	0.85156	0.82731	0.879	0.85318	0.82139	0.7988	0.87308	0.8476
6	0.87316	0.85183	0.89709	0.87458	0.8402	0.82042	0.88838	0.8663

Figure 2. Generalizability coefficients (G or $E\rho^2$) and Dependability indices (Φ) for different number of ratings per essay in various rating designs based on the re-rated (design 1-3) and preliminary data (design 4) for a 3-task test scenario

(a) G-coefficient



(b) Phi- Coefficient

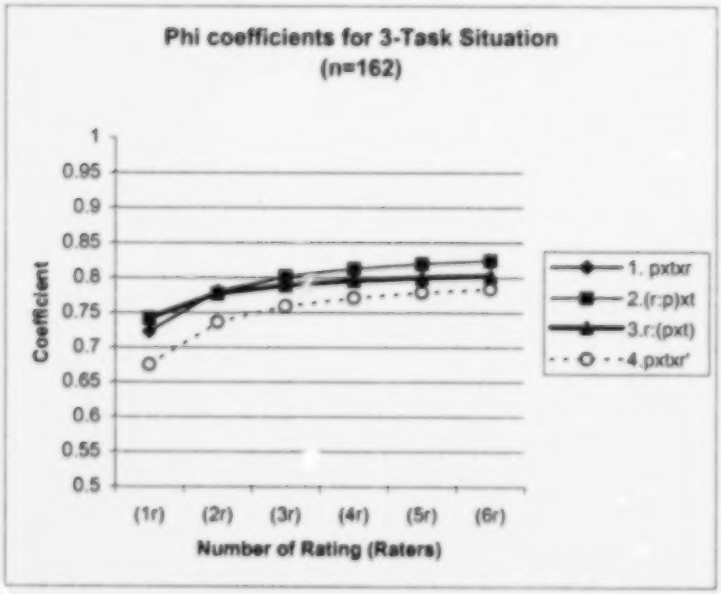


Figure 3. Comparisons of G and Φ coefficients estimated from the initial (p x T x R') and the re-rated (R:(pxT)) data for one and two ratings per essay scenarios for subgroup 3 (n_p=162)

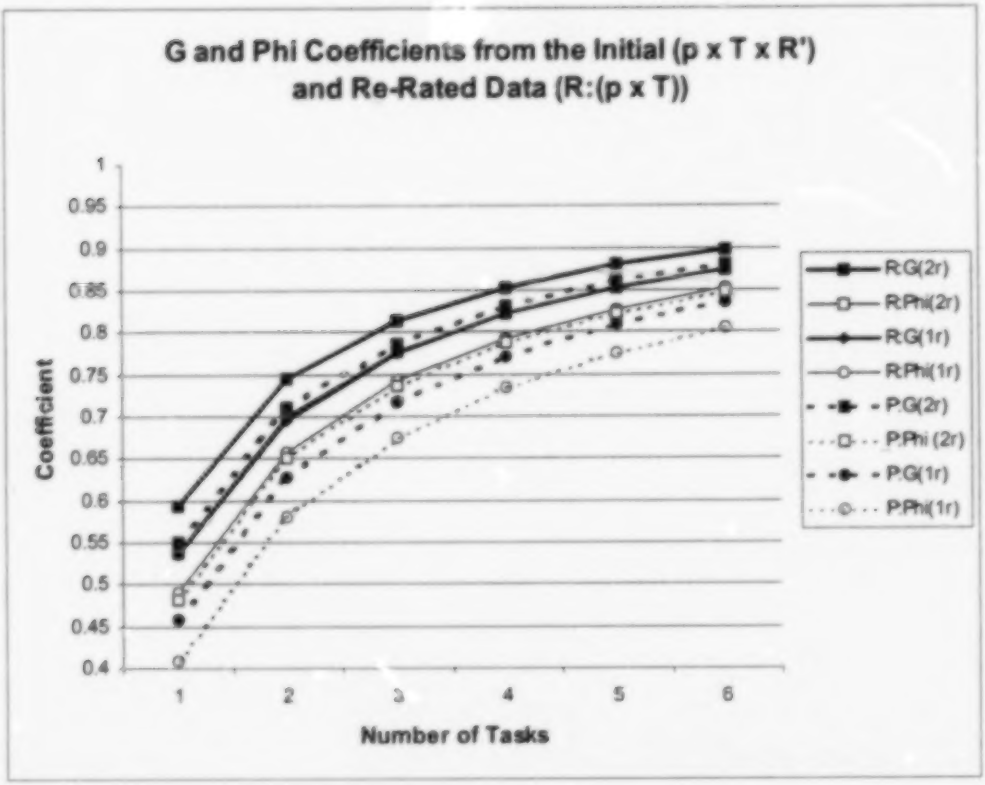


Table 4. Estimated Variance components from G- and D studies for the univariate analysis for Speaking (n=261)

(a) Estimated Variance Components

Effects	G-Study (p x t x r')		D-Study (*p* x T x R')			
	Single Observation		5 Task, 1 Rating		12 Tasks, 2 Rating	
	Variance	Percent	Variance	Percent	Variance	Percent
Person (*p*)	0.66856	51.33%	0.66856	82.11%	0.66856	93.36%
Task (T)	0.02184	1.68%	0.00437	0.54%	0.00182	0.25%
Rating (R')	0	0.00%	0	0.00%	0	0.00%
Person-by-Task (*p*T)	0.22484	17.26%	0.04497	5.52%	0.01874	2.62%
Person-by-Rating (*p*R')	0.02367	1.82%	0.02367	2.91%	0.01184	1.65%
Task-by-Rating (TR')	0.00342	0.26%	0.00068	0.08%	0.00014	0.02%
Person-by-Task-by Rating (*p*TR')	0.36007	27.65%	0.07201	8.84%	0.015	2.09%
Total	1.3024	100.00	0.81426	100.00	0.7161	100.00%

(b) Estimated G- and Phi coefficients

No. of Tasks	*p* x T x R' Design				
	One Rating Per Speech Sample		Two Ratings Per Speech Sample		Two Ratings Per Sample
	G	Φ	G	Φ	α
1	0.5235	0.5133	0.6160	0.6030	0.6269
2	0.6790	0.6704	0.7573	0.7473	0.7707
3	0.7536	0.7465	0.8200	0.8121	0.8345
4	0.7974	0.7914	0.8554	0.8490	0.8705
5	0.8262	0.8211	0.8781	0.8727	0.8937
6	0.8466	0.8421	0.8940	0.8893	0.9098
7	0.8618	0.8578	0.9056	0.9015	0.9217
8	0.8735	0.8700	0.9146	0.9109	0.9308
9	0.8829	0.8797	0.9217	0.9184	0.9380
10	0.8906	0.8876	0.9274	0.9244	0.9438
11	0.8969	0.8942	0.9322	0.9294	0.9487
12	0.9023	0.8997	0.9362	0.9336	0.9528

Figure 5. Internal consistency reliabilities (α) for averaged ratings, *generalizability coefficients* (G) and *dependability indices* (Φ) for one- and two rating per essay scenarios for different test lengths for Speaking

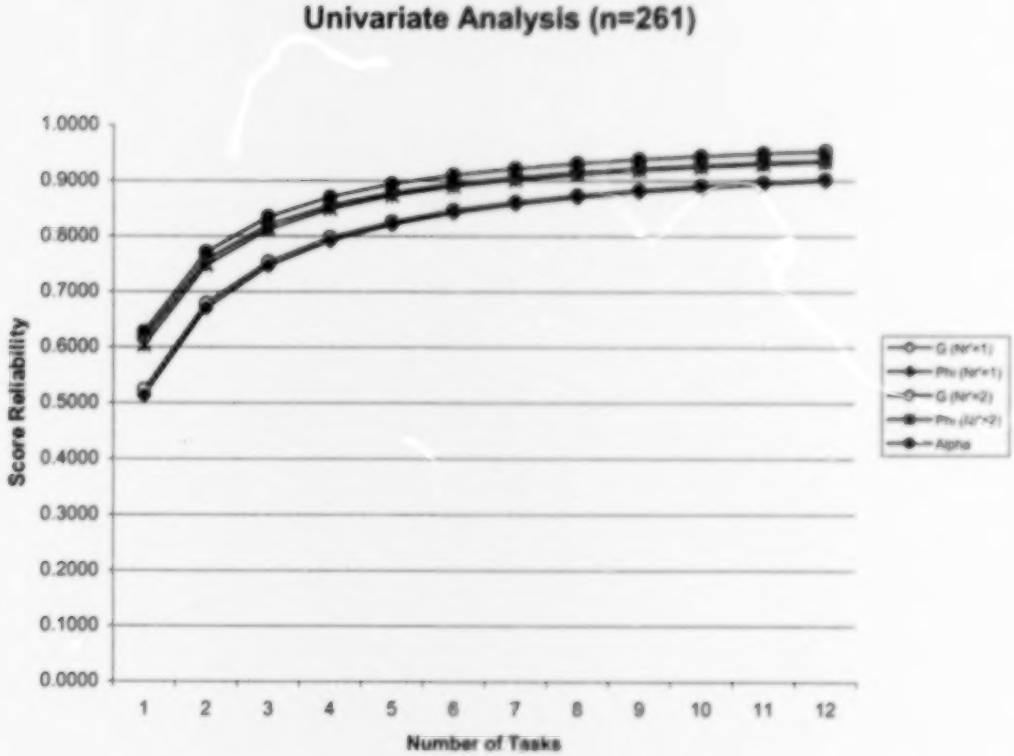


Table 5. *Generalizability coefficients* (G) and *dependability indices* (Φ) for composite scores for different combinations of subsection lengths for fixed total test lengths computed through *multivariate analysis* (n=261).

(a) Estimated Variance Components in the G-study

Effects	G-Study (p x t x r')					
	LS		RS		IS	
	Variance	Percent	Variance	Percent	Variance	Percent
Person (*p*)	0.84204	56.28%	0.77311	55.70%	0.58011	53.36%
Task (T)	0.02124	1.42%	0.00501	0.36%	0.00284	0.26%
Rating (R')	0.00339	0.23%	0.0091	0.66%	0.00061	0.06%
Person-by-Task (*p*T)	0.20864	13.94%	0.15782	11.37%	0.17245	15.86%
Person-by-Rating (*p*R')	0.03843	2.57%	0.16331	11.77%	0.09039	8.32%
Task-by-Rating (TR')	0	0.00%	0.00181	0.13%	0	0.00%
Person-by-Task-by Rating (*p*TR')	0.38246	25.56%	0.27788	20.02%	0.24067	22.14%
Total	1.4962	100.00%	1.38804	100.00%	1.08707	100.00%

(b) Estimated G- and Φ Coefficients in the D-study

D-Study Scenarios				$n_{r'}$ =1		$n_{r'}$ =2	
Total No. of Tasks	No. of LS Tasks	No. of RS Tasks	No. of IS Tasks	G	Φ	G	Φ
3	1	1	1	0.78293	0.78027	0.84642	0.84322
4	2	1	1	0.82657	0.82368	0.87950	0.87620
4	1	2	1	0.80969	0.80733	0.87115	0.86865
4	1	1	2	0.81230	0.81050	0.87014	0.86796
5	3	1	1	0.85305	0.84999	0.89949	0.89625
5	2	2	1	0.84415	0.84220	0.89531	0.89301
5	2	1	2	0.84585	0.84356	0.89450	0.89198
5	1	2	2	0.83294	0.83148	0.88821	0.88653
5	1	1	3	0.82821	0.82686	0.88388	0.88226
5	1	3	1	0.82074	0.81803	0.88284	0.88043
6	2	1	3	0.85654	0.85464	0.90361	0.90159
6	1	1	4	0.83774	0.83663	0.89265	0.89136
11	4	2	5	0.90202	0.90079	0.93749	0.93623
12	4	3	5	0.90559	0.90485	0.94055	0.93963

Figure 6. *Generalizability coefficients* (G) and *dependability indices* (Φ) for composite scores for one- and two rating per essay scenarios for different combinations of subsection lengths for fixed total test lengths.